EXPRESS MAIL LABEL NO. EM295147931US
Date of Deposit: September 29, 1998

JRM:jab 9/29/98

PATENT
Attorney's Matter No. 4239-50781/WDN/JRM

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BOX PROVISIONAL APPLICATION
ASSISTANT COMMISSIONER FOR PATENTS
Washington, D.C. 20231

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(b) (2).

TITLE: RATIO-BASED DECISIONS AND THE QUANTITATIVE ANALYSIS OF cDNA MICRO-ARRAY IMAGES

Inventor(s)/Applicant(s):

Chen	Yidong		Rockville, MC
Dougherty	Edward	R.	College Station, TX
Bittner	Michael	L.	Rockville, MD

Enclosed are:

[x] 25 pages of specification.
[x] 6 sheet(s) of drawings.
[x] Exhibit A, article from *Journal of Biomedical Optics*, October, 1997.
[ ] Verified Statement(s) (Declaration) Claiming Small Entity Status:
- [ ] Independent Inventor (37 CFR 1.9(f) and 1.27(b)).
- [ ] Small Business Concern (37CFR 1.9(f) and 1.27(c)).
- [ ] Non-profit Organization (37CFR 1.9(f) and 1.27(d)).
- [ ] Non-Inventor Supporting a Claim by Another for Small Entity Status (37 CFR 1.9(c) and 1.27(b)).

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

[x] Yes, the name of the U.S. Government agency and the Government contract number are: National Institutes of Health, Department of Health and Human Services

EXPRESS MAIL LABEL NO. EM295147931US
Date of Deposit: September 29, 1998

JRM:jab 9/29/98

PATENT
Attorney's Matter No. 4239-50781/WDN/JRM

Provisional Filing Fee Amount:

[x] $150, large entity
[ ] $ 75, small entity

[X] The Commissioner is hereby authorized to charge any additional fees which may be required in connection with the filing of this provisional application and recording any assignment filed herewith, or credit over-payment, to Account No. 02-4550. A copy of this sheet is enclosed.

Address all telephone calls to Joel Meyer at telephone number (503) 226-7391.

Address all correspondence to:

KLARQUIST SPARKMAN CAMPBELL
LEIGH & WHINSTON, LLP
One World Trade Center, Suite 1600
121 S.W. Salmon Street
Portland, OR 97204

Respectfully submitted,

KLARQUIST SPARKMAN CAMPBELL
LEIGH & WHINSTON, LLP

Date: September 29, 1998

By _____
Joel R. Meyer
Registration No. 37,677

One World Trade Center, Suite 1600
121 S.W. Salmon Street
Portland, Oregon 97204
Telephone (503) 226-7391
Facsimile (503) 228-9446

cc: John Fahner-Vihtelic, NIH No. E-212-98/0
Yidong Chen, Ph.D.
William D. Noonan, M.D.
Docketing Secretary

RATIO-BASED DECISIONS AND THE QUANTITATIVE ANALYSIS OF cDNA MICRO-ARRAY IMAGES

FIELD OF THE INVENTION

The invention relates to quantitative analysis of gene expression in cDNA micro-array images.

BACKGROUND OF THE INVENTION

The recent development of complementary DNA micro-array technology provides a powerful analytical tool for human genetic research (M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270(5235), 467-70, 1995). One of its basic applications is to quantitatively analyze fluorescence signals that represent the relative abundance of mRNA from two distinct tissue samples. cDNA micro-arrays are prepared by automatically printing thousands of cDNAs in an array format on glass microscope slides, which provide gene-specific hybridization targets. Two different samples (of mRNA) can be labeled with different fluors and then co-hybridized on to each arrayed gene. Ratios of gene-expression levels between the samples are calculated and used to detect meaningfully different expression levels between the samples for a given gene.

Biological Background and cDNA Micro-array Technology

A cell relies on its protein components for a wide variety of its functions. The production of energy, the biosynthesis of all component macromolecules, the maintenance of cellular architecture and the ability to act upon intra and extracellular stimuli are all protein dependent. Each cell within an organism contains the information necessary to produce the entire repertoire of proteins which that organism can specify. This information is stored as genes within the organism's DNA genome. The number of human genes is estimated to be 30,000 to 100,000. Within any individual cell, only a portion of the possible gene set is present as protein. Some of

the proteins present in a cell are likely to be present in all cells. These proteins serve functions required in every type of cell, and can be thought of as "housekeeping" proteins. Other proteins serve specialized functions only required in particular cell types. For example, muscle cells contain specialized proteins that form the dense contractile fibers of a muscle. Given that a large part of a cell's specific functionality is determined by the genes it is expressing, it is logical that transcription, the first step in the process of converting the genetic information stored in an organism's genome into protein, would be highly regulated by the control network that coordinates and directs cellular activity.

Regulation is readily observed in studies that scrutinize activities evident in cells configuring themselves for a particular function (specialization into a muscle cell) or state (active multiplication or quiescence). As cells alter their status, coordinate transcription of the protein sets required for this state can be observed. As a window both on cell status and on the system controlling the cell, detailed, global knowledge of the transcriptional state could provide a broad spectrum of information useful to biologists. Knowledge of when and in what types of cell the protein product of a gene of unknown function is expressed would provide useful clues as to the likely function of that gene. Determination of gene-expression patterns in normal cells could provide detailed knowledge of the way in which the control system achieves the highly coordinated activation and deactivation required for development and differentiation of a mature organism from a single fertilized egg. Comparison of gene expression patterns in normal and pathological cells could provide useful diagnostic "fingerprints" and help identify aberrant functions which would be reasonable targets for therapeutic intervention.

The ability to carry out studies in which the transcriptional state of a large number of genes is determined has, until recently, been severely inhibited by limitations on our ability to survey cells for the presence and abundance of a large number gene transcripts in a single experiment. A primary limitation has been the small number of identified genes. In the case of humans, only a few thousand of the complete set (30,000 to 100,000 genes) has been physically purified and characterized

to any extent. Another significant limitation has been the cumbersome nature of transcription analysis. Even a large experiment on human cells would track expression of only a dozen genes, clearly an inadequate sampling for inference about so complex a control system.

Two recent technological advances have provided the means to overcome some of these limitations to examining the patterns and relationships in gene transcription. The cloning of molecules derived from mRNA transcripts in particular tissues, followed by application of high throughput sequencing to the DNA ends of the members of these libraries has yielded a catalog of expressed sequence tags (ESTs) (M.S. Boguski and G.D. Schuler, "ESTablishing a human transcript map," *Nature Genetics,* 10(4), 369-71, 1995). These signature sequences provide unambiguous identifiers for a large cohort of genes. At present, approximately 40,000 human genes have been "tagged" by this route, and many have been mapped to their genomic location (G.D. Schuler and M.S. Boguski, et al., "A gene map of the human genome," *Science*, 274(5287), 540-6, 1996).

Additionally, the clones from which these sequences were derived provide analytical reagents which can be used in the quantitation of transcripts from biological samples. The nucleic acid polymers, DNA and RNA, are biologically synthesized in a copying reaction in which one polymer serves as a template for the synthesis of an opposing strand which is termed its complement. Even after separation from each other, these strands can be induced to pair quite specifically with each other to form a very tight molecular complex, a process called *hybridization.* This specific binding is the basis of most analytical procedures for quantitating the presence of a particular species of nucleic acid, such as the mRNA specifying a particular protein gene product. Micro-array technology, a recent hybridization-based process that allows simultaneous quantitation of many nucleic acid species, has been described (M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270(5235), 467-70, 1995; J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, and J.M. Trent, "Use of a cDNA microarray to analyse gene expression patterns in human

cancer," *Nature Genetics*, 14(4), 457-60 ("DeRisi"), 1996; M. Schena, D. Shalon, R. Heller, A. Chai, P.O. Brown, and R.W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. U.S.A.*, 93(20), 10614-9, 1996). This technique combines robotic spotting of small amounts of individual, pure nucleic acid species on a glass surface, hybridization to this array with multiple fluorescently labeled nucleic acids, and detection and quantitation of the resulting fluor tagged hybrids with a scanning confocal microscope. When used to detect transcripts, a particular RNA transcript (an mRNA) is copied into DNA (a cDNA) and this copied form of the transcript is immobilized on a glass surface. The entire complement of transcript mRNAs present in a particular cell type is extracted from cells and then a fluor-tagged cDNA representation of the extracted mRNAs is made in vitro by an enzymatic reaction termed reverse-transcription. Fluor-tagged representations of mRNA from several cell types, each tagged with a fluor emitting a different color light, are hybridized to the array of cDNAs and then fluorescence at the site of each immobilized cDNA is quantitated.

The various characteristics of this analytic scheme make it particularly useful for directly comparing the abundance of mRNAs present in two cell types. Visual inspection of such a comparison is sufficient to find genes where there is a very large differential rate of expression. A more thorough study of the changes in expression requires the ability to discern more subtle changes in expression level and the ability to determine whether observed differences are the result of random variation or whether they are likely to be meaningful changes.

SUMMARY OF THE INVENTION

The invention provides a method for analyzing expression ratios to determine significant differences in sample expressions across the gene population discernible on a micro-array. The method assumes sample expression levels are independent, levels are normally distributed, and there is a constant coefficient of variation for the entire gene set (a biochemical consequence of the mechanics of transcript production). Using these assumptions, the method derives the probability distribution of the ratio, finds the

maximum-likelihood estimator for the distribution, and employs an iterative procedure for signal calibration.

With this approach, a computer-implemented method can process a single image and identify outliers. Our implementation of this method measures expression levels in digitized micro-array images. In a preprocessing phase, a non-parametric statistical technique extracts cDNA sites on the slide. The method then analyzes the expression ratio using a confidence interval and hypothesis test to quantify the significance of differences in expression ratio.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram illustrating a micro-array system used to compare the abundance of mRNAs present in two cell types.

Fig. 2 is an example of a cDNA micro-array image created using the system shown in Fig. 1.

Fig. 3 is a diagram illustrating an example of a target patch, a target mask, and a target site in a cDNA micro-array image.

Figs. 4A and 4B illustrate target detection results at different significant levels.

Fig. 5 is a graph of ratio density functions for c = 0.05, 0.1, and 0.2, where c represents the coefficient of variation of the ratio density functions.

Fig. 6 is a graph illustrating upper and lower limits for different confidence levels.

Fig. 7 is a graph illustrating ratio density functions for m = 0.5, 1, and 2 when c = 0.1, where m represents the gain factor relating the mean values of the red and green signals.

Fig. 8 is a flow diagram illustrating a method for analyzing expression ratios.

Fig. 9 is a block diagram illustrating a computer system which serves as an operating environment for an implementation of the invention.

DETAILED DESCRIPTION

Capturing the cDNA Micro-array Image

Figure 1 is a diagram illustrating a system for capturing a cDNA micro-array image. This particular system was used in an experiment to compare the abundance of mRNAs present in two cell types. In this experiment, an array of cDNAs was hybridized with a green fluor-tagged representation of mRNAs extracted from a tumorigenic melanoma cell line (UACC-903) and a red fluor-tagged representation of mRNAs from a non-tumorigenic derivative of the original cell line (UACC-903 +6). Monochrome images of the fluorescent intensity observed for each of the fluors are then combined by placing each image in the appropriate color channel of an RGB image, as shown in Fig. 2. In this composite image, one can visualize the differential expression of genes in the two cell lines. Intense red fluorescence at a spot indicates a high level of expression of that gene in the non-tumorigenic cell line with little expression of the same gene in the tumorigenic parent. Conversely, intense green fluorescence at spot indicates high expression of that gene in the tumorigenic line, with little expression in the non-tumorigenic daughter line. When both cell lines express a gene at similar levels, the observed array spot is yellow.

The experiment illustrated above represents only one example of an application of the ratio-based method for analyzing expression levels. In addition to color intensity values, it is also possible to use other ways of labeling probes and then measuring expression levels at target sites. Examples of other labeling techniques include chemifluorescent and radioactive labeling. It is also possible to analyze expression levels of more than two probes in the ratio-based method by selecting one of the probes as a reference.

The following section describes a technique for segmenting target sites in a micro-array image. This technique identifies pixels within a target patch that form a target site. The objective is to identify pixel locations where hybridization has occurred, and to distinguish such locations from the background and noise. These pixel

locations can then be used to estimate expression levels more accurately at the site. While the segmentation is adapted for analyzing color intensity values of an image, it can also be adapted to other signal intensity values. Other segmentation methods may be used in the alternative.

Image Processing and Mann-Whitney Segmentation

Assuming DNA products from two samples have equal probability to hybridize to the target, intensity measurement is a function of the quantity of the specific DNA products available within each sample. Locally (or pixelwise), the intensity measurement is also a function of the concentration of the target segments. On the scanning side, the fluorescent light intensity also depends on the power and wavelength of the laser, quantum efficiency of photo-multiplier tube and efficiency of other electronic devices. The resolution of a scanned image is largely determined by processing requirements and the acquisition speed. The scanning stage imposes a calibration requirement, though it may be relaxed later. The image analysis task is to extract the average fluorescence intensity from each target site (cDNA region).

There are several fluorescent light sources from each slide: background, target, the target hybridized with sample 1 or sample 2, and (possibly) glass surface. The average intensity within a target site is measured by the median image value on the site. This intensity serves as a measure of the total fluors emitted from the sample mRNA probes hybridized on the target site. The median is used as the average to mitigate the effect of outlying pixel values created due to noise.

Some image processing is required prior to intensity measurement. Most is quite standard and need not be described here. For instance, the image needs to be segmented into target patches but this task is straightforward since the robot positions the cDNA targets in a predetermined manner. Because the number of pixels in the target site is limited, both smoothing and sharpening filters need to be avoided.

The difficult image processing task is to identify the target site within the target patch (see Fig. 3). Each target site is somewhat annular owing to how the robot finger places the cDNA on the slide and how the slide is treated; however, there is variability

in this placement (within the patch) from image to image and from target to target. This variability can be so great that the target region is simply a collection of subregions within the nominal circular target region. This instability in the target region is manifested in the irregular way the mRNA is hybridized to the target and the consequent irregular brightness pattern (created by the fluors) within the target site. It is important that mRNA intensity be measured over these fluor regions because only they correspond to probe-hybridized-to-target areas. Conventional adaptive thresholding segmentation techniques are unsatisfactory when the signal is weak because there is no marked transition between foreground and background. Standard morphological methods also fail because for weak signals there is no consistent shape information for the target area.

To overcome these difficulties we propose a pixel selection method based on the Mann-Whitney test. There are three key points associated with the proposed Mann-Whitney approach: (1) it associates a confidence level to every intensity measurement based on the significance level of the test and, if desired, it enables multiple readouts at different confidence levels, (2) it meets the real-time requirement of the system, and (3) it is a distribution-free test, thereby eliminating the need for a normality assumptions.

We briefly describe the Mann-Whitney test as employed herein. Suppose X_1, $X_2,\ldots, X_n$ and Y_1, $Y_2,\ldots, Y_m$ are independent samples arising from two random variables X and Y possessing means μ_X and μ_Y, respectively. The rank-sum statistic W, which is the sum of the ranks of all X samples in the combined ordered sequence of the X and Y samples, is used to test the null hypothesis,

$$H_0: \ \mu_X - \mu_Y = 0$$
$$H_1: \ \mu_X - \mu_Y > 0 \qquad (1)$$

The Mann-Whitney criterion reveals the relation between the positions of the X and Y positions in the combined ordered sequence. Rejection of H_0 occurs when $W \geq w_{\alpha,n,m}$, the critical value corresponding to the significance level α.

A target site is segmented from the target patch according to the following procedure. A predefined target mask is used to identify a portion of the target patch

that contains the target site. The target mask is based on the geometry of the potential target area and can be constructed from specially tagged targets or other strong targets (e.g., the target mask is obtained by finding all strong targets, aligning them together, averaging and then thresholding). We randomly pick 8 sample pixels from the known background (outside the target mask) as $Y_1, Y_2, \ldots Y_8$, and select the lowest 8 samples from within the target mask as $X_1, X_2, \ldots X_8$. The rank-sum statistic W is calculated and, for a given significance level α, compared to $w_{\alpha,8,8}$. We choose 8 samples here for both foreground and background because the Mann-Whitney statistic is approximately normal when $m = n \geq 8$. If the null hypothesis is not rejected, then we discard some predetermined number (perhaps only 1) of the 8 samples from the potential target region and select the lowest 8 remaining samples from the region. The Mann-Whitney test is repeated until the null hypothesis is rejected. When H_0 is rejected, the target site is taken to be the 8 pixels causing the rejection together with all pixels in the target mask whose values are greater than or equal to the minimum value of the eight. The resulting site is said to be a target site of significance level α. If the null hypothesis is never rejected, then it is concluded that there is no appreciable probe at the target site. Furthermore, one can require that the Mann-Whitney target site contain at minimum some number of pixels for the target site to be considered valid and measured for fluor intensity. Figure 4a and 4b show the detection results of target sites at $\alpha = 0.0001$ and $\alpha = 0.05$, respectively, where detected site boundaries are superimposed with original images. Once a target site is determined, gene expression is measured by the median of the target site minus the median of the background area (outside the target mask area).

After segmenting the target site from the image data, the ratio of the expression levels at the site are computed by computing a ratio of the red intensity average and the green intensity average for pixel locations in the target site. The average of the red and green signals may be computed as the mean or median of the respective intensity values for each color. The next phase then analyzes the ratio distribution of the target sites.

Probability Density Function of Ratio Parameters

Having extracted the target site, we now use another computer program that uses the expression ratio to determine whether or not gene expression differs significantly for the red and green sample. Equal distributions for red and green values lead to a red/green ratio close to 1 and significantly unequal distributions lead to a red/green ratio significantly different from 1. In examining expression ratios, two points need to be taken into consideration. First, even if red and green measurements are identically distributed, the mean of the ratio distribution will not be 1; second, the hypothesis test needs to be performed on expression levels from a single micro-array. A salient factor in using expression ratios rather than expression differences is that gene expression levels are determined by intrinsic properties of each gene, which means that expression-level differences vary widely between genes regardless of the truth of the null hypothesis. Therefore, it is inappropriate to pool gene-expression difference statistics across the micro-array. Labeling the red and green micro-array values for the genes by $R_1, R_2,\ldots, R_n$ and $G_1, G_2,\ldots, G_n$, respectively, the desired hypothesis test is

$$H_0: \mu_{R_k} = \mu_{G_k}$$
$$H_1: \mu_{R_k} \neq \mu_{G_k} \qquad (2)$$

using the test statistic $T_k = R_k/G_k$. This requires finding a critical region for T_k, recognizing that the mean of T_k under the null hypothesis is not 1.

It is well-known that working with ratio distributions can be problematic and recent research on the matter is generally confined to normality study of the ratio distribution, and numerical calculations. However, as we now discuss, a special situation arises for gene expression that permits a more detailed statistical analysis, as well as hypothesis tests and confidence intervals based on a single micro-array.

While it would be possible to gather data on the routine level of expression for each specific gene in each specific tissue, this would be a very difficult undertaking. The method currently requires substantial quantities of mRNA (and thus tissue) for each determination. Extending the studies to pathological situations would further complicate the ability to gather material for replicates, since it will initially be

necessary to assume diseases with complex molecular etiologies may have many forms, making pooling of samples from different individuals counterproductive. The most practical and informative version of an assay of this type would be achieved if information on the variance of all or most of the genes in a sample could be used to derive a statistically sound measurement of variance for each individual transcript. Fortunately, it appears that the biology of transcription makes such an approach possible.

A transcript's abundance at a given time is governed by the current rates of production and degradation of that transcript. As would be expected of a system faced with routine generation and destruction of these information intermediates, the processes which produce and destroy transcripts rely on common, core enzymatic machinery (polymerases and nucleases) whose specificity of activity is modulated by accessory proteins that bind to the core enzymes, the nucleic acid sites of action or both. As might also be expected of a system that must constantly synthesize and hydrolyze tens of thousands of molecules, molecular interactions are based on very similar intermolecular affinities. Nimbleness at this scale requires that the core machinery operate without too much bias, so that no single or small class of transcripts consumes too large a share of the machinery's capacity. This type of bulk processing is thus predicted to be an approximation of a much simpler reaction, in which the level of a transcript will depend roughly on the concentration of the accessory factors driving its selection, and the variations for any particular transcript would be expected to be normally distributed and constant (as a fraction of abundance) relative to most of the other transcripts. Such assumptions on the variances produce a special situation that can be exploited to great advantage, allowing the use of the variation data from all transcripts surveyed to be pooled to estimate the global variation of transcript synthesis and destruction. An important caveat to this hypothesis is that transcripts present at extremely high or extremely low levels could require a different method of control of synthesis/degradation and would not necessarily have variances representative of transcripts present at a common level.

Assuming there is constant coefficient of variation c for the entire gene set,

$$\begin{aligned} \sigma_{R_k} &= c\mu_{R_k} \\ \sigma_{G_k} &= c\mu_{G_k} \end{aligned} \qquad (3)$$

Under the null hypothesis H_0, $\mu_{R_k} = \mu_{G_k}$. Letting μ_k denote the common value, the condition of Eq. 3 becomes $\sigma_{R_k} = \sigma_{G_k} = c\mu_k$. From the experimental protocol, we assume that R_k and G_k are independent, identically distributed normal random variables.

If X and Y are continuous random variables, $T = X/Y$, and X and Y possess the joint probability density function $f_{X,Y}(x, y)$, then, the probability distribution function for T is

$$\begin{aligned} F_T(t) &= P(X \le tY, Y > 0) + P(X \ge tY, Y < 0) \\ &= \int_0^\infty \left[\int_{-\infty}^{ty} f_{X,Y}(x, y)dx \right] dy + \int_{-\infty}^0 \left[\int_{ty}^{\infty} f_{X,Y}(x, y)dx \right] dy \end{aligned} \qquad (4)$$

For X and Y independent, differentiation yields the probability density function for T as

$$\begin{aligned} f_T(t) &= \int_0^\infty y f_{X,Y}(ty, y)dy - \int_{-\infty}^0 y f_{X,Y}(ty, y)dy \\ &= \int_0^\infty y f_X(ty)\, f_Y(y)dy - \int_{-\infty}^0 y f_X(ty)\, f_Y(y)dy \end{aligned} \qquad (5)$$

where the second equality follows from independence.

We apply Eq. 5 under the normality, independence, and constant-coefficient-of-variation conditions. Since micro-array intensity measurements are positive, densities for both red and green values are assumed to be 0 for negative arguments. The error created by the simultaneous normality and positive-value assumptions is negligible because measurement intensities are sufficiently positive to render the portions of the left distribution tails falling to the left of the y axis negligible. Letting $T_k = R_k/G_k$,

$$f_{T_k}(t) = \int_0^\infty g f_{R_k}(tg)\, f_{G_k}(g)dg - \int_{-\infty}^0 g f_{R_k}(tg)\, f_{G_k}(g)dg$$
$$= \int_0^\infty \frac{1}{\sigma_{R_k}\sqrt{2\pi}} e^{-\frac{(tg-\mu_{R_k})^2}{2\sigma_{R_k}^2}} \frac{1}{\sigma_{G_k}\sqrt{2\pi}} e^{-\frac{(g-\mu_{G_k})^2}{2\sigma_{G_k}^2}} g dg \qquad (6)$$
$$= \frac{1}{2\pi c^2} \int_0^\infty e^{-\frac{(tu-1)^2}{2c^2}} e^{\frac{(u-1)^2}{2c^2}} u du$$

where the second equality follows from the positive-value assumption and the third from Eq. 3 and the substitution $g/\mu_k = u$. Note that the density for T_k is independent of k. This property is not merely a consequence of Eq. 3, but depends on normality.

The integration of Eq. 6 yields a solution that is given by the standard error equation. Notice that the second exponential in the integrand is similar to the normal density function with $\mu = 1$ and $\sigma = c$. When c is small (less than 0.3), the second exponential is close to 0 for $u < 0$. Therefore, by extending the integration to $-\infty$, we have the approximation

$$f_{T_k}(t) \approx \frac{1}{2\pi c^2} \int_{-\infty}^\infty e^{-\frac{(tu-1)^2}{2c^2}} e^{\frac{(u-1)^2}{2c^2}} u du$$
$$= \frac{(1+t)\sqrt{1+t^2}}{c(1+t^2)^2\sqrt{2\pi}} e^{-\frac{(t-1)^2}{2c^2(1+t^2)}} \qquad (7)$$

The approximation error of Eq. 7 can be numerically evaluated. For example, given $c = 0.3$, at $t = 1.0$ the approximation error between Eqs. 6 and 7 is 4.8×10^{-8}, and at $t = 3.0$ the error is 1.2×10^{-8}. Figure 5 depicts the probability density function given in Eq. 7 for $c = 0.05$, 0.1, and 0.2. The density function of Eq. 7 is an asymmetric function and its peak is close to 1, under the null hypothesis. Since Eqs. 6 and 7 are not functions of k, we denote the density function by $f_T(t; c)$ with parameter c.

Confidence Intervals and Maximum-Likelihood Estimation

Confidence intervals can be obtained via Eq. 7. Table 1 lists the upper (right) limit and lower (left) limit of 95% confidence intervals for different c values, as well as the means and standard deviations of the corresponding distributions. As functions of

c, the confidence-interval limits, mean and standard deviation can be approximated by polynomial functions

$$y = a_3c^3 + a_2c^2 + a_1c + a_0 \qquad (8)$$

Table 2 gives the appropriate polynomial coefficients for the upper limit, lower limit, mean and standard deviation. Figure 6 provides curves for 95%, 90%, 85% and 80% confidence levels. Most results obtained here have been verified by Monte Carlo simulation. Referring back to the hypothesis test of Eq. 1, for each k, the acceptance region for the test statistic T_k is the confidence interval for the appropriate value of c and the confidence level.

Typically, c needs to be estimated from the data. Using the density of Eq. 7 we can obtain a maximum-likelihood estimator for c. The likelihood function is

$$L(c) = \prod_{i=1}^{n} \frac{(1+t_i)\sqrt{1+t_i^2}}{c(1+t_i^2)^2\sqrt{2\pi}} e^{-\frac{(t_i-1)^2}{2c^2(1+t_i^2)}} \qquad (9)$$

where $t_1, t_2, \ldots, t_n$ are ratio samples taken from a single collection of expression values, for example, all ratios from the housekeeping genes in a micro-array. The maximum-likelihood criterion requires that $d[\log L(c)]/dc = 0$. Hence, the estimator for c is

$$\hat{c} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{(t_i-1)^2}{(1+t_i^2)}} \qquad (10)$$

Uncalibrated Signals

The null hypothesis of equal means is appropriate for calibrated signal acquisition but in practice this may not be the case. Therefore we consider the uncalibrated situation in which the means of the red and green signals are related by a constant amplification (or reduction) gain factor m, $\mu_{R_k} = m\mu_{G_k}$. If $m \geq 1$, then the red signal is stronger than the green. We can follow the same derivation as in the calibrated case except that now the ratio density has two parameters, c and m. This results in the recursive relation

$$f_T(t;c,m) = \frac{1}{m} f_T(t/m, c, 1) \quad (11)$$

where $f_T(\bullet; c,1)$ is given by Eq. 7. Figure 7 shows cases for $m = 0.5$, 1, and 2 (when $c = 0.1$). For $m = 0.5$ we expect R_k/G_k to be about 0.5, which is what Fig. 7 indicates.

In the uncalibrated setting, estimators are required for both c and m; however, a closed-form solution as in the calibrated case is precluded by reliance on the recursion of Eq. 11. Our program proceeds iteratively to obtain estimators. As shown in Table 1, the means for different c values are very close to 1 when $m = 1$.

Input Dist.	Output Distribution Parameters					
c.v. (c)	L. Limit	U. Limit	mean (μ)	dev. (σ)	c.v. (σ/μ)	peak s_{max}
0.01	0.972	1.026	1.0001	0.014	0.0141	0.9998
0.02	0.945	1.052	1.0004	0.028	0.0283	0.9992
0.03	0.919	1.08	1.0009	0.042	0.0424	0.9982
0.04	0.894	1.108	1.0016	0.056	0.0567	0.9968
0.05	0.869	1.137	1.0025	0.071	0.0710	0.9950
0.06	0.845	1.167	1.0036	0.085	0.0854	0.9928
0.07	0.822	1.198	1.0049	0.100	0.0998	0.9903
0.08	0.798	1.230	1.0065	0.115	0.1144	0.9873
0.09	0.776	1.263	1.0083	0.130	0.1291	0.9840
0.10	0.754	1.297	1.0103	0.145	0.1440	0.9804
0.11	0.732	1.332	1.0125	0.161	0.1591	0.9764
0.12	0.710	1.369	1.0150	0.177	0.1744	0.9720
0.13	0.689	1.407	1.0178	0.193	0.1900	0.9673
0.14	0.669	1.447	1.0208	0.210	0.2059	0.9623
0.15	0.648	1.488	1.0242	0.227	0.2222	0.9570
0.16	0.628	1.531	1.0278	0.245	0.2389	0.9514
0.17	0.609	1.576	1.0318	0.264	0.2561	0.9455
0.18	0.589	1.623	1.0362	0.283	0.2740	0.9393
0.19	0.570	1.672	1.0409	0.304	0.2925	0.9329
0.20	0.551	1.724	1.0460	0.326	0.3118	0.9262
0.21	0.532	1.778	1.0515	0.349	0.3319	0.9192
0.22	0.514	1.835	1.0574	0.372	0.3526	0.9122
0.23	0.495	1.895	1.0637	0.397	0.3740	0.9049
0.24	0.477	1.958	1.0702	0.423	0.3958	0.8974
0.25	0.459	2.026	1.0770	0.450	0.4178	0.8898
0.26	0.441	2.098	1.0840	0.477	0.4400	0.8820
0.27	0.424	2.174	1.0909	0.504	0.462	0.8740
0.28	0.407	2.257	1.0979	0.531	0.4840	0.8660
0.29	0.390	2.346	1.1047	0.558	0.5055	0.8579
0.30	0.373	2.442	1.1113	0.585	0.5266	0.8497

Table 1. Lower and upper limits at 95% confidence level, and other statistics of ratio density.

Intuitively, when two signals are approximately the same, the mode of the ratio density will be around 1. Therefore, a usual calibration practice is to move the ratio histogram mode to 1 when the red and green signals are not calibrated. This calibration procedure is not strictly correct because the peak of the ratio density changes with parameter *c*. To account for this effect, we first assume the population mean μ_0 to be 1 and let the first approximation m_1 of the calibration parameter be the sample mean. The sample

data is then calibrated by m_1. After that, Eq. 10 is used to estimate the first approximation c_1 of c. Estimation proceeds by iteratively repeating the procedure.

Figure 8 is a flow diagram illustrating the steps performed by a computer program to compute ratios of expression levels. The method employed by this program proceeds as described below. The following description refers to steps in the flow diagram shown in Fig. 8 by reference numbers.

1. Initialize mean estimate $\hat{\mu}_0$ of the ratio density of Eq. 7 to be 1 (equivalent to assuming $c_0 = 0$). See step 800.
2. Calibrate ratio samples so that the input red and green signals are approximately equal by taking the estimator of m, say $\hat{m}_i$, to be the sample mean divided by the previous mean estimator,

$$\hat{m}_i = \frac{1}{\hat{\mu}_{i-1}}\left(\frac{1}{n}\sum_{j=1}^{n} t_j\right) \qquad (12)$$ See step 802.

The calibration factor is taken to be $1/\hat{m}_i$. The normalized ratio data set is

$$(t'_1, t'_2, \ldots, t'_n) = (t_1 / \hat{m}_i, t_2 / \hat{m}_i, \ldots, t_n / \hat{m}_i) \quad (13)$$ See step 804.

3. Use the maximum-likelihood estimator of Eq. 10 to calculate $\hat{c}_i$ by evaluating the estimator with the newly calibrated ratio data (t_i', $i = 1, 2, \ldots, n$). See step 806.
4. Estimate the mean $\hat{\mu}_i$ of the ratio distribution, given the new $\hat{c}_i$, by using the polynomial regression given in Table 2 ($\mu = 0.364c^3 + 1.279c^2 - 0.0427c + 1.001$). See step 808.
5. Repeat steps 2 through 4 until a satisfactory result is obtained. Figure 8 illustrates this iterative process with a loop from decision block 810 back to step 802. Since the ratio mean μ is close to 1 for even relatively large values of c, five iterations are usually sufficient.
6. Upper and lower confidence limits (θ_1, θ_2) can be obtained using Table 1 or 2 (Step 812), and then converting them to the desired interval ($\theta_1 \cdot \hat{m}$, $\theta_2 \cdot \hat{m}$)(Step 814).

Conf. Level		a_3	a_2	a_1	a_0	goodness of fit (R^2)
	Lower limit	-2.805	2.911	-2.706	0.979	0.999994

95%	Upper limit	28.644	-2.830	3.082	0.989	0.99993
	Lower limit	-5.002	4.462	-3.496	0.9968	0.99998
99%	Upper limit	78.349	-15.161	4.810	0.9648	0.99998
	mean (μ)	0.364	1.279	-0.0427	1.001	0.9997
	Std. Dev.	6.259	0.190	1.341	0.0022	0.9998

Table 2. Parameters of fitting polynomial functions.

To verify the accuracy of the iterative method under the H_0 condition ($\mu R_k = \mu G_k$), we performed the following simulation assuming 100 red and 100 green intensity data. For $k = 1, 2, \ldots, 100$, the kth red signal's (representing the kth gene expression level in sample 1) mean intensity μ_{R_k} is drawn from a uniform random process with range from 100 to 30,000 (simulating 16-bit integer range). For a given m and c value, along with the normality for both red and green signals, we generate a single datum for both the kth red and green signal, thereby obtaining a sample of the red/green ratio for each k. Simulations were done for m from 0.3 to 3 with a step of 0.1, and for c from 0.01 to 0.3 with a step of 0.01, each simulation involving the full iterative procedure. The entire simulation was repeated 30 times for each value of m and c. Average estimation errors for c and m are under 1% where error for c is defined as $|(\hat{c} - c)/c|$ and for m similarly.

Experimental Results

Consider the superimposed micro-array image from UACC-903 (red channel) and UACC-903(+6) (green channel) shown in Fig. 2. The full array contains 1,368 clone segments. A total of 88 ratio samples whose ratios are believed to be about 1 (whose gene expression levels are assumed unchanged in both cell lines), such as housekeeping genes, are listed in Web site http://www.nhgri.nih.gov/DIR/LCG/ARRAY/expn.html. Since acquisition does not insure perfect calibration, the iterative procedure is used. The result is as follows:

m 1.1316

c 0.1727 (or 17.27%)

99% confidence interval: (0.566, 1.977)

The step by step illustration of the iterative estimation is listed in Table 3. The 99.0%

confidence interval for $c = 0.1727$ and $m = 1.1316$ is (0.56617, 1.97684).

Step i	Sample Scaling factor	c_i (Eq. 10)	μ_i (Table 2)	m_i (Eq. 12)
initial	—	—	$\mu_0 = 1.0$	$m_0 = 1.1697$
1	$1/m_0$	0.1741	1.03425	1.1420
2	$1/m_1$	0.1728	1.03370	1.1315
3	$1/m_2$	0.1727	1.03365	1.1316
4	$1/m_3$	0.1727	Stop!	—

Table 3. Step by step illustration of the iterative estimation.

Based on this interval, 92 ratio samples are found to be significant. Of these, 70 were found to be significant using the inappropriately narrow confidence interval reported by DeRisi (see Web site http://www.nhgri.nih.gov/DIR/LCG/ARRAY /expn.html).

Gene Name	R/G Ratio
pre-mRNA splicing factor SRp7	2.33
casein kinase I delta	2.33
MAC25	2.32
endothelin-1 (EDN1)	2.30
B12 protein	2.25
RSU-1/RSP-1	2.25
Id1	2.24
similar to induced myeloid leukemia cell differentiation protein	2.22
male-enhanced antigen mRNA (Mea)	2.20
PP15 (placental protein 15)	2.20
vascular endothelial GF	2.18
calphobindin II	2.18
similar to mouse transplantation antigen p35B	2.15
22kDa smooth muscle protein (SM22)	2.15
alternative guanine nucleotide-binding regulatory protein (G)	2.13
nuclear autoantigen GS2NA	2.13
cadherin-associated protein-related (cap-r)	2.13
mitochondrial phosphate carrier protein	2.12
alpha NAC	2.10
thymopoietin beta	2.08
B lymphocyte serine/threonine protein kinase	2.07
platelet alpha SNAP	2.06
lamin B2 (LAMB2)	2.06
CMAR	2.06
inosine-5'-monophosphate dehydrogenase (IMP)	2.02
I-Rel	1.99
DNA-binding protein (CROC-1A)	1.99

Gene Name	R/G Ratio
polyA binding protein	1.98
bcr (break point cluster gene)	0.57
mitotic feedback control protein Madp2 homolog	0.55
protein-tyrosine phosphatase	0.55
Human poly(ADP-ribose) synthetase	0.53

Table 4. Named genes, shown in decreasing ratio order, are additional genes found to have different expression levels in a chromosome 6 suppressed melanoma cell line than in its tumorigenic parent (99% confidence level). The original findings were reported by DeRisi.

Table 4 lists the ones missed by the confidence limits reported by DeRisi. Some of the newly found significant changes are biologically interesting and further bolster general impressions resulting from the original cohort of genes showing significant changes. Two further examples of the tendency of the chromosome 6 suppressed line toward increased expression of genes associated with differentiation are the myeloid leukemia cell differentiation protein (mcl1) and the cell adhesion regulator protein (CAR/CMAR). Increased expression of the mcl1 gene has been found to be a very early indicator of induced differentiation in cancer cells. See A. Umezawa, T. Maruyama, et al, "Induction of mcl1/EAT, Bcl-2 related gene, by retinoic acid or heat shock in the human embryonal carcinoma cells, NCR-G3," Cell Structure Function, 21(2), 132-50, (1996); and T. Yang, H. L. Buchan, et al. "MCL-1, a member of the BLC-2 family, is induced rapidly in response to signals for cell differentiation or death, but not to signals for cell proliferation," J. Cell Physiology, 166(3), 523-36 (1996).

Increased expression of the CAR gene has been correlated with reduced spontaneous metastatic potential in the HT-29 (human adenocarcinoma) cell line, presumably due to a greater repertoire of integrins with increased adherence of the cells to the extracellular matrix. See H. Yamamoto, F. Itoh, et al. "Inverse association of cell adhesion regulator messenger RNA expression with metastasis in human colorectal cancer," Cancer Research, 56(15), 3605-9, (1996).

In addition to the tendency toward expression of genes associated with differentiation, changes are observed that suggest that the suppressed cells are more capable of modulating oncogene activity. In addition to the strong increase in p21 expression previously seen, significant increase in the expression of the ras suppressor

Rsu-1 is observed. Rsu-1 has been shown to be a potent inhibitor of Jun kinase activation. L. Masuelli and M. L. Cutler "Increased expression of the Ras suppressor Rsu-1 enhances Erk-2 activation and inhibits Jun kinase activation," Molecular Cell Biology , 16(10), 5466-76, (1996).

Operating Environment for the Invention

Figure 9 and the following discussion are intended to provide a brief, general description of a suitable computing environment for the computer programs described above. The method for analyzing expression ratios is implemented in computer-executable instructions organized in program modules. The program modules include the routines, programs, objects, components, and data structures that perform the tasks and implement the data types described above.

While Fig. 9 shows a typical configuration of a desktop computer, the invention may be implemented in other computer system configurations, including multiprocessor systems, microprocessor-based or programmable consumer electronics, minicomputers, mainframe computers, and the like. The invention may also be used in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

The computer system shown in Fig. 9 includes a personal computer 920, including a processing unit 921, a system memory 922, and a system bus 923 that interconnects various system components including the system memory to the processing unit 921. The system bus may comprise any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using a bus architecture such as PCI, VESA, Microchannel (MCA), ISA and EISA, to name a few. The system memory includes read only memory (ROM) 924 and random access memory (RAM) 925. A basic input/output system 926 (BIOS), containing the basic routines that help to transfer information between elements within the personal computer 920, such as during start-up, is stored in ROM 924. The

personal computer 920 further includes a hard disk drive 927, a magnetic disk drive 928, e.g., to read from or write to a removable disk 929, and an optical disk drive 930, e.g., for reading a CD-ROM disk 931 or to read from or write to other optical media. The hard disk drive 927, magnetic disk drive 928, and optical disk drive 930 are connected to the system bus 923 by a hard disk drive interface 932, a magnetic disk drive interface 933, and an optical drive interface 934, respectively. The drives and their associated computer-readable media provide nonvolatile storage of data, data structures, computer-executable instructions (program code such as dynamic link libraries, and executable files), etc. for the personal computer 920. Although the description of computer-readable media above refers to a hard disk, a removable magnetic disk and a CD, it can also include other types of media that are readable by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, and the like.

A number of program modules may be stored in the drives and RAM 925, including an operating system 935, one or more application programs 936, other program modules 937, and program data 938. A user may enter commands and information into the personal computer 920 through a keyboard 940 and pointing device, such as a mouse 942. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 921 through a serial port interface 946 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port or a universal serial bus (USB). A monitor 947 or other type of display device is also connected to the system bus 923 via an interface, such as a display controller or video adapter 948. In addition to the monitor, personal computers typically include other peripheral output devices (not shown), such as speakers and printers.

The personal computer 920 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 949. The remote computer 949 may be a server, a router, a peer device or other common network node, and typically includes many or all of the elements described relative to

the personal computer 920, although only a memory storage device has been illustrated in Figure 9. The logical connections depicted in Figure 9 include a local area network (LAN) 951 and a wide area network (WAN) 952. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the personal computer 920 is connected to the local network 951 through a network interface or adapter 953. When used in a WAN networking environment, the personal computer 920 typically includes a modem 954 or other means for establishing communications over the wide area network 952, such as the Internet. The modem 954, which may be internal or external, is connected to the system bus 923 via the serial port interface 946. In a networked environment, program modules depicted relative to the personal computer 920, or portions thereof, may be stored in the remote memory storage device. The network connections shown are merely examples and other means of establishing a communications link between the computers may be used.

Conclusion

Ratios are used to quantify gene-expression distinctions on a cDNA micro-array arising from different samples. Under the mathematical conditions assumed for average mRNA expression intensities, the ratio distribution has been derived, maximum-likelihood estimation characterized, and calibration achieved via an iterative algorithm. Empirically, a careful mathematical analysis of calibration and confidence limits has revealed significant gene-expression ratios that were missed with a less precise analysis.

While the above description focuses on a specific implementation of a method for quantitatively analyzing gene expression levels, the invention is not limited to this specific implementation. This analysis method generally applies to any form of detectable signals as long as the signals are (a) hybridized to the same site; and (b) proportional to the corresponding input probes. Condition (a) is important because the amount of target cDNA or its binding quality will be reflected in the detected signal intensity. When multiple probes are hybridized to the same target, they are exposed to

the same condition. Condition (b) is important because it avoids competition for hybridization. In practice, the target concentration should be in great excess relative to the probes such that a disproportionate signal favoring more abundant species is eliminated.

More than two probes can be hybridized to the same slide as long as the target cDNA concentration is sufficiently large that no competition for hybridization sites is observed and the signal levels are proportional to the input probe. If one color (corresponding to a cell type) is selected as the reference color, the ratio-based method can be applied directly. For example, the color red could be selected as the reference, and the intensity level for each of the other probes' colors could be divided by the reference intensity level for red to compute the ratio.

The signals need not correspond to color values. In general, the signals captured from the micro-array represent the quantity of gene expression at particular sites. This quantity can be conveyed via other signal types corresponding to some attribute of the tag or label such as a gray scale intensity, a color intensity, radiation level, etc.

In addition, there are a variety of alternative labeling methods to measure expression levels. The mRNA labeling can be achieved using radiation (P33, P32, etc.), fluorescence (Cy3, Cy5, etc.), chemifluorescence (e.g., alkaline phosphate-based chemifluorescence), physical labels (e.g., biotin labels), or any other direct or indirect labeling as long as signal levels representing the expression levels of each probe can be detected separately. One example of an imaging system suitable for capturing signals representing expression levels is the STORM ® image system from Molecular Dynamics, Inc., of Sunnyvale, CA. This system combines PhosphorImager ® system technology with non-radioactive labeling techniques: direct fluorescence and chemifluorescence.

Another aspect of the ratio-based method that may vary depending on the implementation is how and which genes are selected as the internal control genes. Under the null hypothesis, the ratios of every gene expression level possess an identical distribution, regardless of the actual expression levels. However, in practice, not all

genes satisfy the null hypothesis. The experiments described above use a set of house-keeping genes, having expression levels that are assumed unchanged in both test samples, for ratio parameter estimation, including the normalization procedure. While this appears to be a preferred approach, it is not necessary to use such a set of control genes in all implementations of the ratio-based method.

There are drawbacks, however, to using a very small sample of genes (e.g., one or two genes selected as a standard with a known ratio) or including every gene in the sample. In the case of a small sample, the entire array will be erroneously calibrated if the standard genes are not correctly detected. In the inclusive approach, every ratio is used without any selection. The result will be biased if a significant fraction of test samples have changed expression levels relative to the reference sample.

A carefully selected internal control gene set with about 100 genes provides a robust calibration result, where the population distribution is estimated even though some of the genes in the pre-selected set may not satisfy the null hypothesis. The confidence interval derived from the distribution can then be applied to the entire array. In other words, the calibration procedure is based on the behavior of a carefully selected set of "housekeeping" genes, as opposed to one or two standard genes or an inclusive set of genes.

The selection of internal control genes can be categorized into three methods: (a) selection by biological house-keeping function, (b) selection by statistical stability over a large collection of array data, and (c) a combination of (a) and (b).

For more information, please see Yidong Chen, Edward R. Dougherty, Michael L. Bittner, "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Journal of Biomedical Optics,* October, 1997, attached as Appendix A, which is hereby incorporated by reference.

In view of the variety of ways in which the invention can be implemented, it is not limited to the specific implementation described above. Instead, the invention includes all that reasonably falls within the scope and spirit of the following claims.

We claim:

1. A method for analyzing gene expression in cDNA micro-array images comprising:

(a) identifying target sites for genes in a cDNA micro-array image;

(b) computing a maximum-likelihood estimator for a coefficient of variation for expression ratio samples, where the expression ratio samples are taken from a collection of expression values for each gene in a set of genes identified in the micro-array image, and each gene in the set has an expression ratio representing a ratio of expression level for a first cell type to an expression level for a second cell type;

(c) computing an estimated mean value of a ratio distribution of the expression ratios;

(d) computing a confidence interval for the ratio distribution; and

(e) using the confidence interval to identify ratio samples for corresponding genes in the micro-array image for which the expression ratio is outside the confidence interval.

2. The method of claim 1 further including:

initializing a mean estimate for the ratio distribution;

calibrating expression ratio samples by computing a scaling factor between the expression levels of the first and second cell types such that the expression level for the first and second cell types are approximately equal after being adjusted by the scaling factor; and

repeating steps b-d until the maximum-likelihood estimator and estimated mean value are within a threshold amount of a previously calculated maximum-likelihood estimator and estimated mean value.

3. The method of claim 2 further including:

capturing the micro-array image for an array of cDNAs hybridized with a first color tagged representation of mRNAs extracted from the first cell type and a second color tagged representation of mRNAs extracted from the second cell type, where color

intensity values of the first color in the micro-array image represent an expression level of the first cell type, and color intensity values of the second color represent an expression level of the second cell type; and

computing an expression ratio for each gene in the set as a ratio of an average of the first color intensity values to an average of the second color intensity values for pixel locations in the identified target site for the gene.

RATIO-BASED DECISIONS AND THE QUANTITATIVE ANALYSIS OF cDNA MICRO-ARRAY IMAGES

ABSTRACT OF THE DISCLOSURE

Gene expression can be quantitatively analyzed by hybridizing fluor-tagged mRNA to targets on a cDNA micro-array. Comparison of gene expression levels arising from co-hybridized samples is achieved by taking ratios of average expression levels for individual genes. In an image-processing phase, a method of image segmentation identifies cDNA target sites in a cDNA micro-array image. The resulting cDNA target sites are analyzed based on a hypothesis test and confidence interval to quantify the significance of observed differences in expression ratios. In particular, the probability density of the ratio and the maximum-likelihood estimator for the distribution are derived, and an iterative procedure for signal calibration is developed.

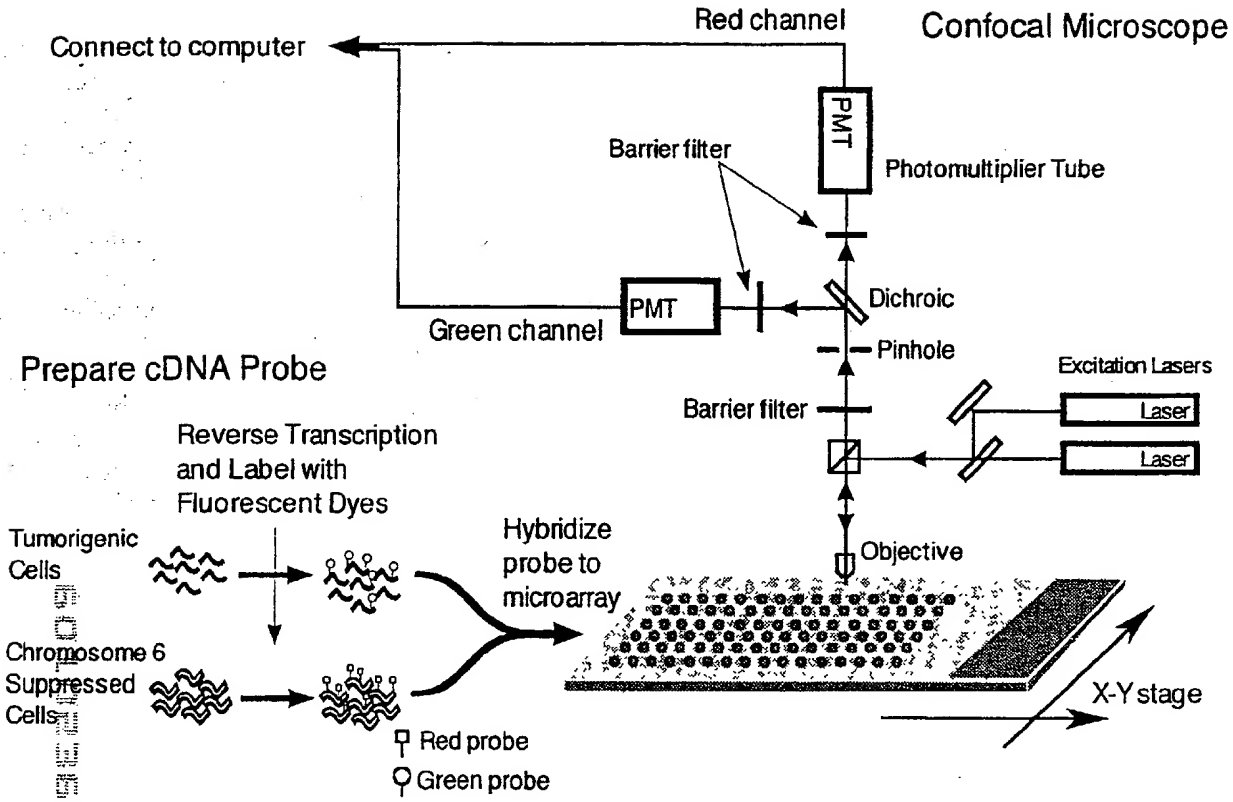


Fig. 1 Illustration of a microarray system.

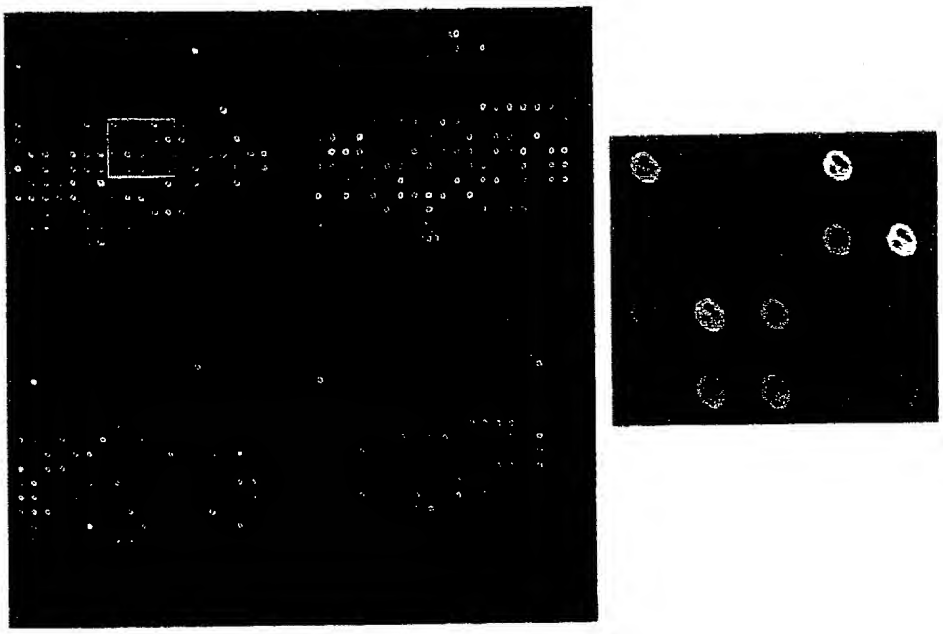


Fig. 2 cDNA microarray image.

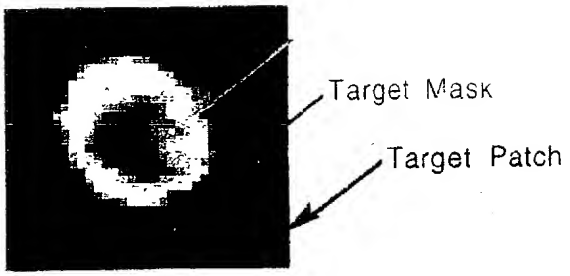
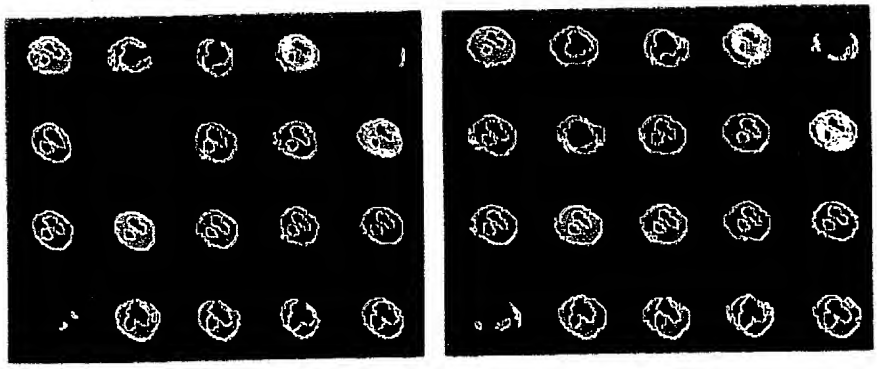


Fig. 3 Target patch, mask, and site.



(a) Detection result at $\alpha = 0.0001$. (b) Detection result at $\alpha = 0.05$.

Fig. 4 Target detection results at different significant levels.

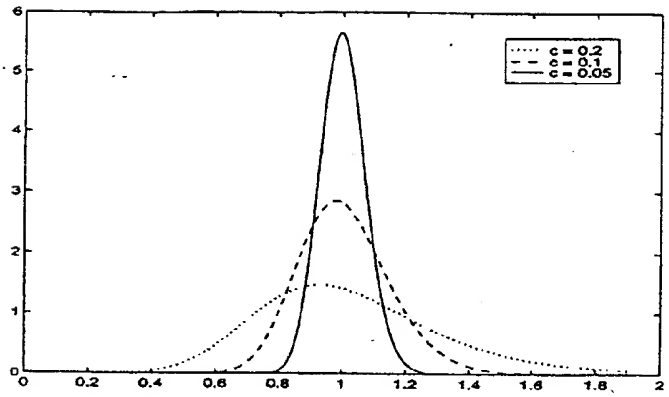


Fig. 5 Ratio density functions for c=0.05, 0.1, and 0.2.

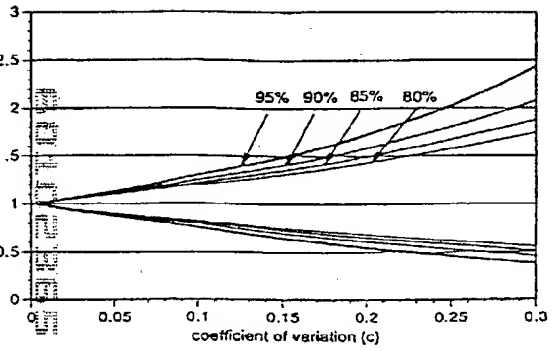


Fig. 6 Limits for different confidence levels.

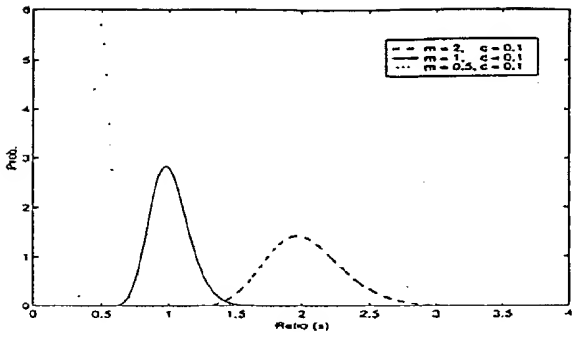


Fig. 7 Ratio density functions for $m=0.5$, 1, and 2 when $c=0.1$.

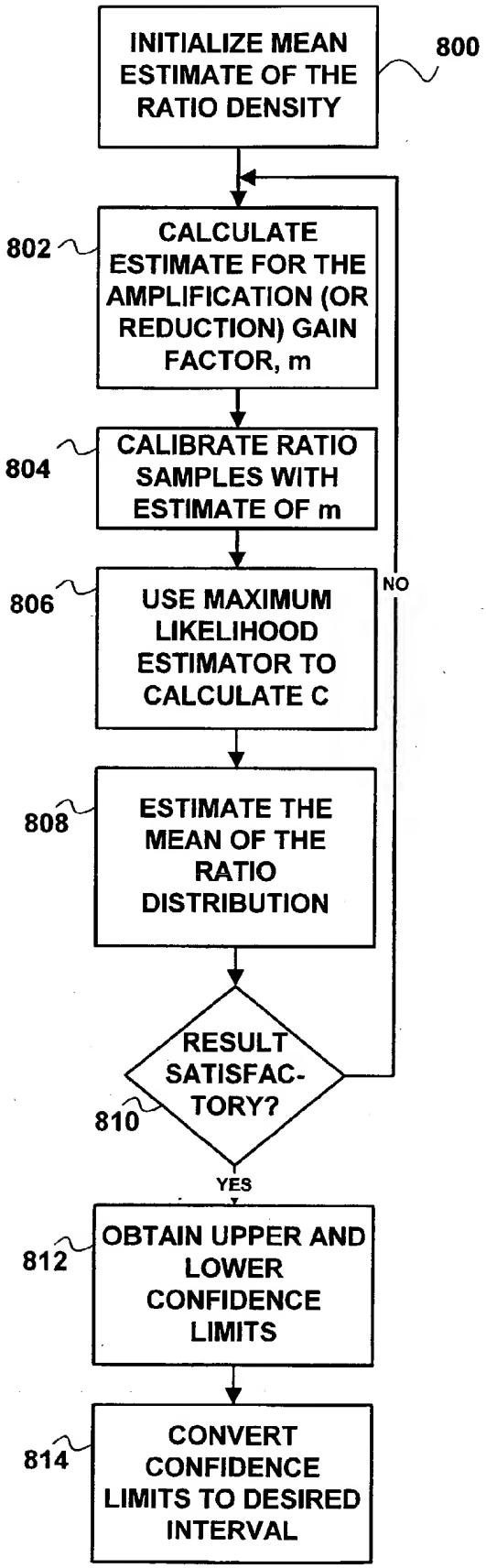
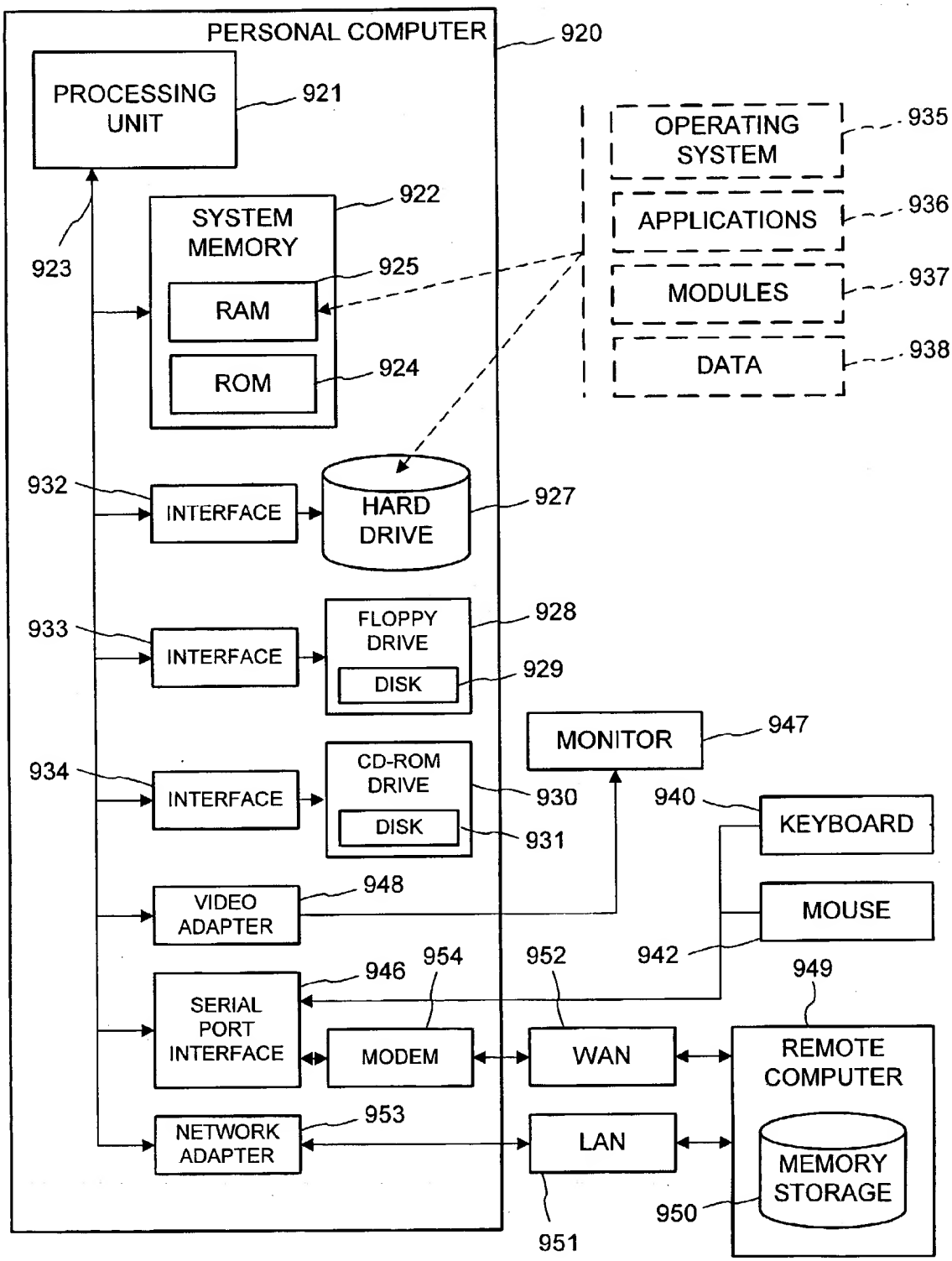

INITIALIZE MEAN ESTIMATE OF THE RATIO DENSITY
800
802
CALCULATE ESTIMATE FOR THE AMPLIFICATION (OR REDUCTION) GAIN FACTOR, m
804
CALIBRATE RATIO SAMPLES WITH ESTIMATE OF m
806
USE MAXIMUM LIKELIHOOD ESTIMATOR TO CALCULATE C
808
ESTIMATE THE MEAN OF THE RATIO DISTRIBUTION
RESULT SATISFAC-TORY?
810
NO
YES
812
OBTAIN UPPER AND LOWER CONFIDENCE LIMITS
814
CONVERT CONFIDENCE LIMITS TO DESIRED INTERVAL

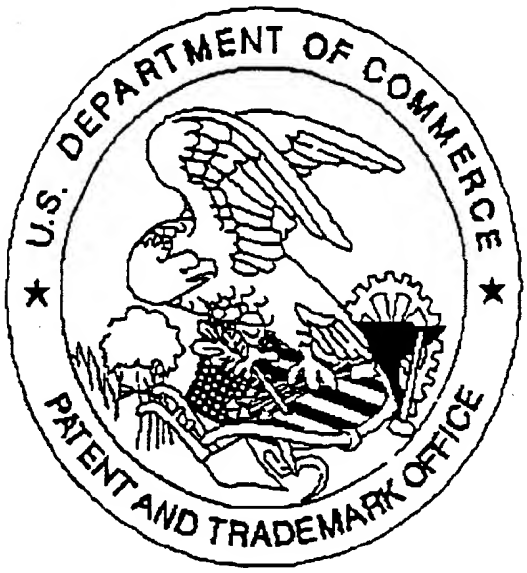
FIG. 8

FIG. 9


PERSONAL COMPUTER
920
PROCESSING UNIT
921
923
SYSTEM MEMORY
922
RAM
925
ROM
924
OPERATING SYSTEM
935
APPLICATIONS
936
MODULES
937
DATA
938
932
INTERFACE
HARD DRIVE
927
933
INTERFACE
FLOPPY DRIVE
928
DISK
929
934
INTERFACE
CD-ROM DRIVE
930
DISK
931
MONITOR
947
940
KEYBOARD
MOUSE
942
948
VIDEO ADAPTER
946
954
952
949
SERIAL PORT INTERFACE
MODEM
WAN
REMOTE COMPUTER
953
NETWORK ADAPTER
LAN
MEMORY STORAGE
950
951

United States Patent & Trademark Office

Office of Initial Patent Examination – Scanning Division



Application deficiencies were found during scanning:

☐ Page(s)_____ of Declarations _____ were not present for scanning.
(Document title)

☐ Page(s)_____ of _____ were not present for scanning.
(Document title)

☐ Scanned copy is best available.